Ilham Wahyu Adli

1301173380

Dataset

Dataset didapatkan dari salah satu lomba yang diadakan oleh Institut Pertanian Bogor yang dinamakan Statistika Ria dan Festifal Sains Data 2020. Dataset yang disajikan pada perlombaan tersebut merupakan klasifikasi teks apakah sebuah narasi merupakan hoax/fakta, data narasi dan pelabelan tersebut didapatkan dari Turn Back Hoax. Berikut adalah bentuk data yang diberikan:

```
dataset = pd.read_excel('Data Latih BDC.xlsx')
dataset.head()
```

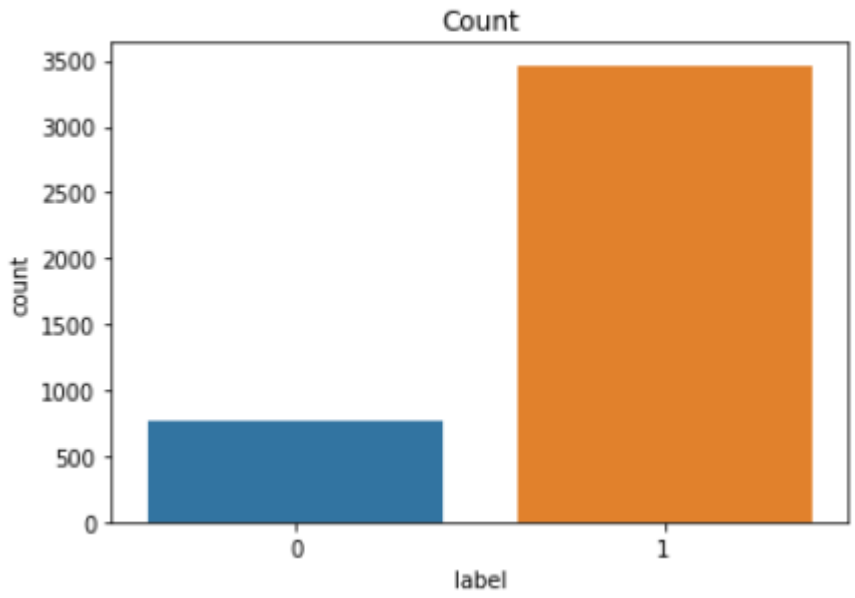
	ID	label	tanggal	judul	narasi	nama file gambar
0	71	1	2020-08-17 00:00:00	Pemakaian Masker Menyebabkan Penyakit Legionna...	A caller to a radio talk show recently shared ...	71.jpg
1	461	1	2020-07-17 00:00:00	Instruksi Gubernur Jateng tentang penilangan ...	Yth.Seluruh Anggota Grup Sesuai Instruksi Gube...	461.png
2	495	1	2020-07-13 00:00:00	Foto Jim Rohn: Jokowi adalah presiden terbaik ...	Jokowi adalah presiden terbaik dlm sejarah ban...	495.png
3	550	1	2020-07-08 00:00:00	ini bukan politik, tapi kenyataan Pak Jokowi b...	Maaf Mas2 dan Mbak2, ini bukan politik, tapi k...	550.png
4	681	1	2020-06-24 00:00:00	Foto Kadrun kalo lihat foto ini panas dingin	Kadrun kalo lihat foto ini panas dingin . .	681.jpg

Gambar 1 Dataset yang akan digunakan, dengan label 1 berarti hoax dan 0 adalah fakta.

Dalam pengerjaan tugas ini, saya memilih atribut narasi dan judul sebagai penentu apakah teks tersebut termasuk kedalam klasifikasi 1 atau 0. Kedua atribut tersebut saya gabungkan dan dimasukan kedalam kolom narasi.

Eksplorasi Dataset

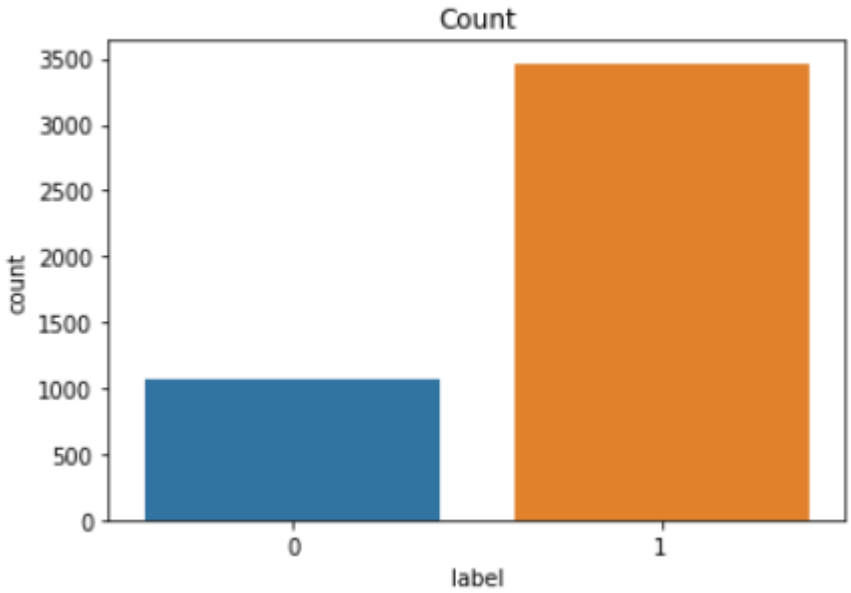
```
Counter({1: 3465, 0: 766})
```



Gambar 2 Jumlah Label 0 dan 1 dalam dataset

Berdasarkan jumlah dataset pada gambar 2, dimana label 0 terdapat 766 data dan label 1 terdapat 3465 data hal ini dapat menyebabkan *imbalance class.* Untuk mengatasi masalah ini saya melakukan duplikasi terhadap dataset pada label 0 sebanyak 300 data.

```
Counter({1: 3465, 0: 1066})
```



Gambar 3 Dataset yang telah di-resampling sebanyak 300 data

Hasil Implementasi Algoritma SVM

```
presisi  : 93.33 %
recall   : 77.31 %
f1 score : 82.08 %
```

Gambar 4 Hasil implementasi dengan membagi data uji sebanyak 10% dari dataset

Berdasarkan hasil yang telah ditampilkan diatas algoritma ini termasuk baik untuk melakukan klasifikasi selanjutnya ketika ada teks yang nantinya akan dilakukan klasifikasi.

Analisa Hasil Algoritma SVM

Parameter yang digunakan dalam Algoritma SVM adalah kernel = ‘poly’ dan random_state = 1.

Pengaruh persentase data latih dan uji

		Pembagian Data (dalam persen)				
		90-10	80-20	50-50	20-80	10-90
Metode Evaluasi	Presisi	93,33 %	92,77 %	90,04 %	88,49 %	88,98 %
	Recall	77,31 %	71,77 %	63,11 %	56 %	53,47 %
	F1-Score	82,08 %	76,44 %	65,72 %	54,84 %	50,31 %

Berdasarkan tabel diatas dapat dilihat pengaruh pembagian data latih dan data uji dapat mempengaruhi performa dari algoritma yang diimplementasikan. Semakin banyak data yang dipilih menjadi data latih, semakin tinggi pula ketiga hasil metode evaluasi. Hal ini

membuktikan adanya pengaruh pembagian data latih dan data uji dalam dataset yang diimplementasikan pada tugas ini.

Waktu pelatihan yang diperlukan

Ketika menjalankan algoritma SVM waktu yang diperlukan untuk memprediksi 454 data dengan 4077 data latih tidak memerlukan waktu yang lama. Pada mesin saya waktu yang diperlukan tidak sampai dengan 10 detik untuk menginplementasikan dengan pembagian data seperti diatas. Hal yang paling lama dilakukan ketika menjalankan algoritma adalah ketika membersihkan data yaitu berkisar 1 menit.

Gambar 5 Progress Bar Cleaning Data

Hasil Data Uji

	narasi	label	Prediction	Hasil Akhir
2278	virus corona korlap fpi imam yang tangkal viru...	1	1	True
2843	gubernur nyali tanah sumatera ria revolusi lan...	1	1	True
2102	aku ribka tjiptaning mayoritas anak pki gabung...	1	1	True
2874	mall emporium brusan jadi ngecas hp pakai powe...	1	1	True
1683	turn back hoax tukang nyebar hoax laku ketangk...	1	1	True
428	kelas online makan korban siswa bunuh ponsel f...	1	1	True
926	menteri energi jepang bungkuk menit sebagai me...	1	1	True
966	selamat kepada guna setia whatsapp no whatsapp...	1	1	True
1077	bang saya dapat video yang nama sabobase jatuh...	1	1	True
1541	sosok misterius jokowi oligarki sosok misteriu...	1	1	True

Gambar 6 Prediksi Benar

	narasi	label	Prediction	Hasil Akhir
2091	hoax konten tugas bssn kait hoax menkominfo bs...	0	1	False
1992	berita bohong hoaks ujar benci marak media sos...	0	1	False
3157	aplikasi whatsapp potensi disalahgunakan sebar...	0	1	False
461	tahun pdip mui babi panggan halal konsumsi kal...	0	1	False
806	massa tolak seminar pki mana tni polri isu kom...	0	1	False
3509	cipta jalan trans papua jawab eny wijayanti ma...	0	1	False
382	timbang teknis pilih produk jerman timbang kua...	0	1	False
1958	nu dukung lgbt nu dukung lgbt	0	1	False
792	begni obat sakit asam urat kolestrol tekan dar...	0	1	False
2689	data gambar hasil survei nasional mengatasnama...	0	1	False

Gambar 7 Prediksi Salah

Analisa

Data tersebut dapat salah klasifikasi dikarenakan metode yang digunakan adalah dengan cara mencari kemiripan suatu dokumen dengan TF-IDF. Hal ini mengakibatkan sulitnya menentukan apakah suatu kalimat tersebut merupakan hoax atau fakta jika belum ada pembahasan mengenai suatu hal tersebut pada data latih yang diberikan. Sehingga kemungkinan terjadinya kesalahan prediksi masih dapat terjadi. Dan pengaruh banyaknya data latih sangat berpengaruh ketika menggunakan metode TF-IDF dalam mengklasifikasikan suatu teks. Dan juga adanya imbalance class yang telah dibahas pada bagian eksploritasi dataset juga mempengaruhi hasil dari prediksi yang telah dilakukan.